Protein structures sustain evolutionary drift

Burkhard Rost

A protein sequence folds into a unique three-dimensional protein structure. Different sequences, though, can fold into similar structures. How stable is a protein structure with respect to sequence changes? What percentage of the sequence is 'anchor' residues, that is, residues crucial for protein structure and function? Here, answers to these questions are pursued by analyzing large numbers of structurally homologous protein pairs. Most pairs of similar structures have sequence identity as low as expected from randomly related sequences (8–9%). On average, only 3–4% of all residues are 'anchor' residues. The symmetric shape of the distribution at low sequence identity suggests that for most structures, four billion years of evolution was sufficient to reach an equilibrium. The mean identities for convergent (different ancestor) and divergent (same ancestor) evolution of proteins to similar structures are quite close and hence, in most cases, it is difficult to distinguish between the two effects. In particular, low levels of sequence identity appear not to be indicative of convergent evolution.

Address: EMBL, 69012 Heidelberg, Germany.
E-mail: rost@embl-heidelberg.de

Electronic identifier: 1359-0278-002-S0019

Folding & Design 01 Jun 1997, 2:S19–S24

Introduction

Large-scale studies of protein structure evolution can begin

We have a detailed and ever-widening knowledge of the evolution of DNA sequences. But what do we really know about the evolution of protein structure? Until recently, the answer was: not much. The first detailed structures were determined 27 years ago [1,2], and 14 years ago, the database of atomic-resolution protein structures (PDB) contained just 312 structures [3]. Since then, due to advances in determination methods [4], the PDB has grown exponentially; presently, it holds ~5000 entries. A parallel development has occurred in methods for aligning protein structures [5–16]. Applying these automatic methods to the current PDB, we can now begin to analyze the evolution of protein structure.

Stability of structures with respect to sequence changes enables evolutionary drift

It has long been accepted that each protein sequence folds into a unique 3D structure, and that proteins with similar sequences have similar structures [17]. But exactly how similar do two sequences have to be to have similar structures? In other words, how large is the sequence attractor of a protein structure (i.e., the region in sequence space which maps onto the same fold)? The answer is surprisingly large: essentially all protein pairs of known structure with more than 30 out of 100 residues identical pairwise have similar structures [18]. This high robustness of structures with respect to residue exchanges explains partly the robustness of organisms with respect to gene-replication errors, and it allows much scope for variety in evolution. In recent years, many examples of protein pairs have been uncovered which have similar structures at even lower levels of pairwise sequence identity. At first, this was a surprise [19,20]. However, we are now starting to realize that a low level of sequence identity between similar structures is not the exception.

Here, a previous analysis of protein structure evolution [21] is detailed and extended. This analysis is based on all pairs of proteins in the PDB with similar 3D structures. For each pair, the structures were aligned and the sequence identity (pairwise identical residues) in the aligned regions measured. To minimize bias in regions of higher identity, distributions of pairwise sequence identity were compiled for four entire genomes representing all three terrestrial kingdoms: *Haemophilus influenzae* (prokaryote), *Saccharomyces cerevisiae* (eukaryote), *Mycoplasma genitalium* (prokaryote), and *Methanococcus jannaschii* (archean).

Methodology

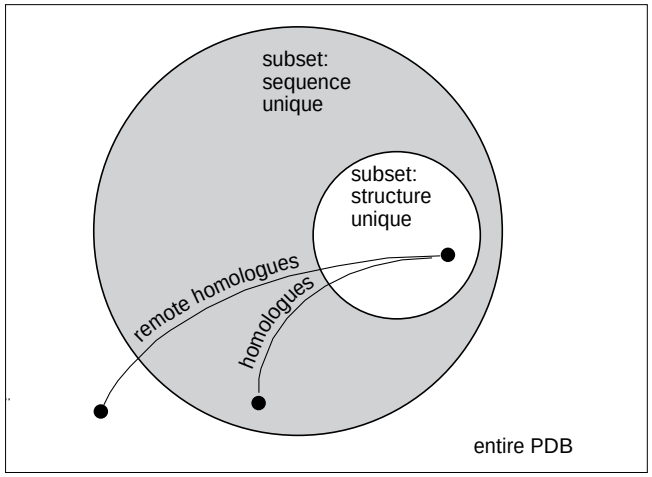
Compiling pairwise sequence identity

The basic score compiled was the level of pairwise sequence identity, i.e., the percentage of residues identical between two aligned proteins. Structure alignments were taken from the FSSP database of structure alignments [22]. The method that produced these alignments (DALI) attempts to superimpose two structures according to their similarity in the pattern of interresidue contacts [23]. Thus, the feature analyzed here (pairwise sequence identity) has not been used by the structural alignment algorithm.

Two regions of pairwise sequence identity: close and remote structural homologues

The level of pairwise sequence identity for which two naturally evolved proteins are guaranteed to have similar structures [17] depends on the alignment length [18]. A 'twilight zone' [24] distinguishes the region in which sequence identity implies structure similarity and the region for which many proteins have different structures. Sander and Schneider [18] defined a length-dependent cut-off line

Figure 1



Avoiding bias by selection of the data set. The analysis presented in this paper was compiled on the basis of the largest possible subsets from the PDB that fulfilled the following criteria: sequence unique – no pair of structures had >25% pairwise sequence identity (according to structural alignment [23]); structure unique – starting from the sequence-unique set, a set was selected in which no pair of structures had a significant structural similarity (defined by the DALI cut-off [23]). This procedure implied the separation into two regions of pairwise sequence identity: (1) remote structural homologues (<25% pairwise sequence identity) – all proteins in the structure-unique set (white inner circle) were aligned against all proteins in the sequence-unique set (grey outer circle); (2) close structural homologues (>25% pairwise sequence identity) – all proteins in the structure-unique set (white inner circle) were aligned against all proteins in the PDB.

which was used in this analysis to separate the region of close structural homologues (pairwise sequence identity >25%) and the region of remote structural homologues (pairwise sequence identity <25%). In the first region, sequence alignment methods produce accurate alignments; in the second region, reliable alignments have to be based on knowledge about structure.

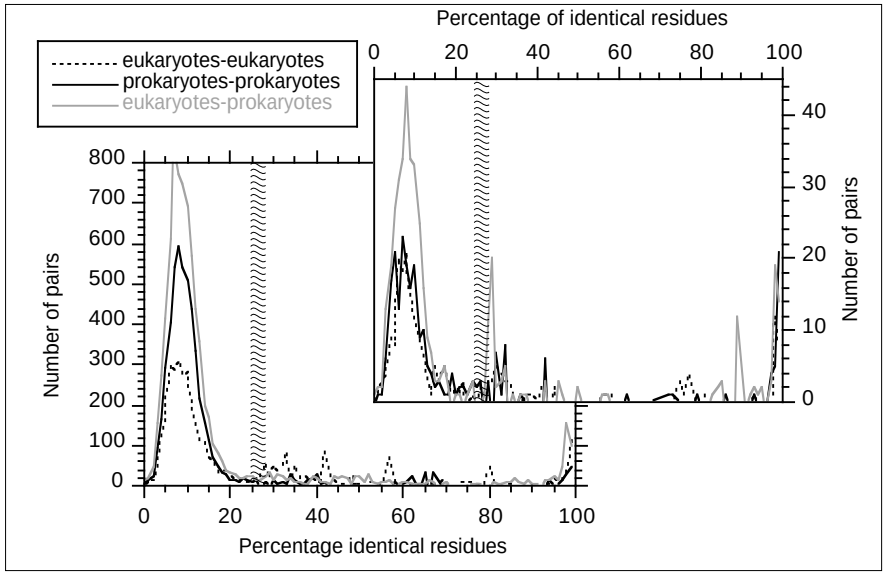
Avoiding bias by populated folds through selection of the data set

First, the largest subset of unique structures was selected (272 proteins for which 148 had at least one remote homologue; Fig. 1; compiled from [22]). Second, the largest subset of sequence-unique structures was selected (849 proteins; Fig. 1; compiled from [22]). To further reduce possible bias, the unique structures were aligned against the set of unique sequences only (instead of against the entire PDB; Fig. 1). The distributions of levels of pairwise sequence identity >25% were generated by aligning all proteins in the 'structure-unique' set against all proteins in the PDB (note that by definition, in the set of sequence-unique structures there is no pair with >25% pairwise sequence identity; Fig. 1). To explore the effect of comparisons between and within the two major terrestrial kingdoms, the alignments were additionally restricted to homologues between: (1) prokaryotes and prokaryotes; (2) eukaryotes and eukaryotes; and (3) mixed pairs, i.e. one protein from prokaryotes, the other from eukaryotes. When starting from the structure-unique set, the counts yielded were too low (upper-right chart in Fig. 2). Therefore, the inter-kingdom and intra-kingdom data were also compiled starting from the sequence-unique subset of the PDB (lower-left chart in Fig. 2).

Exploring four entire genomes

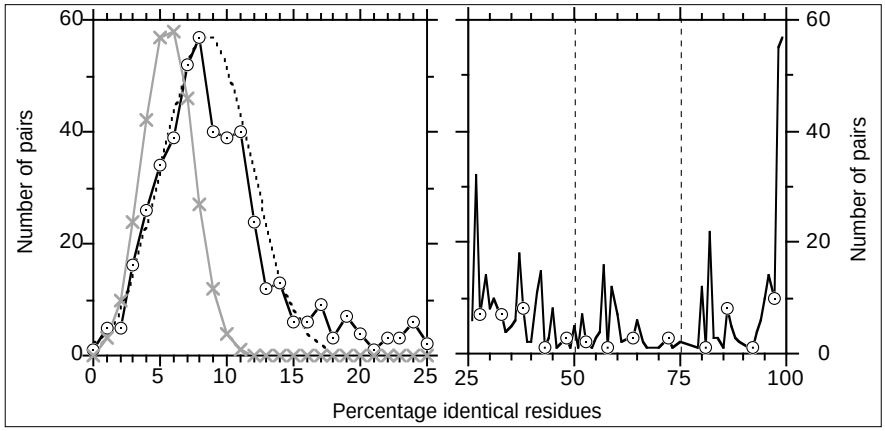
The structure alignments yielded a rather 'noisy' distribution in the region above 40% sequence identity (Fig. 3). A way to less biased distributions has been opened by

Figure 2



Distribution of pairwise sequence identity for structural homologues belonging to the same kingdom. The figure shows the results for all eukaryotes against all eukaryotes (dashed black line), for all prokaryotes against all prokaryotes (solid black line), and for all eukaryotes against all prokaryotes (solid grey line). The lower and upper charts differ in the subsets of the PDB considered (lower-left chart: starting from the sequence-unique subset of the PDB, see Fig. 1; upper-right chart: starting from the structure-unique subset of the PDB, see Fig. 1). The bars indicate that the absolute values of the distribution below and above 25% sequence identity are not comparable. To minimize the effect of database bias below 25%, the start set was aligned against a subset of the PDB in which no two proteins had >25% pairwise sequence identity, whereas above 25%, the start set was aligned against the entire PDB (see Fig. 1).

Figure 3



Distribution of pairwise sequence identity for structural alignments (open circles, black line) and random alignments (left panel only; crosses, grey line). The average sequence identity of all remote structural homologues (<25% pairwise sequence identity, left panel) was ~8.5% (standard deviation 5%). The dashed line is a Gaussian envelope (left panel) fitted to the observed distribution. The average sequence identity of random alignments was ~5.6% (standard deviation 3%).

sequencing the entire genomes of *H. influenzae* (HI) [25], *S. cerevisiae* (YE) [26], *M. genitalium* (MG) [27], and *M. jannaschii* (MJ) [28]. For each genome, distributions were generated by aligning all protein sequences against the SWISS-PROT and the TREMBL [29] databases (using the multiple sequence alignment program MAXHOM [18,30]).

Generating random background distributions

The random background distribution was generated by the following procedure. All proteins in the structure-unique set were randomly superimposed onto all proteins in the sequence-unique set. 'Randomly superimposed' refers to selecting alignment 'begin' and 'end' in both 'aligned' proteins by generating random numbers, i.e. irrespective of the amino acids that were superimposed. A constraint was imposed while randomly selecting pairs: the pairs had to mirror the distribution of alignment length observed by the structurally aligned pairs. This particular construction of random pairs guaranteed that the random background was representative for the set of structurally aligned proteins: as well as singlet frequencies (amino acid composition), higher order correlations (di-, tri-, multipeptide frequencies) were similar. (Therefore, the average value of 5.6% was lower than what would have been expected from superimposition of randomly shuffled sequences.)

Results

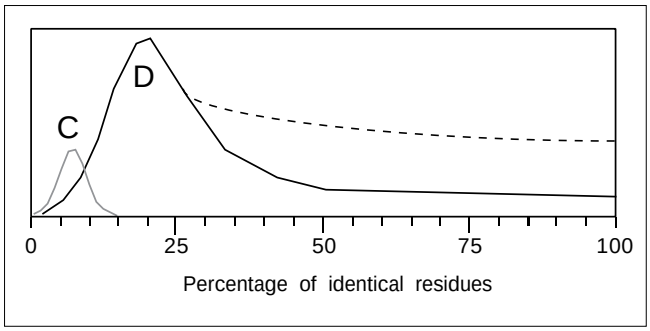
Expected distributions for convergent and divergent evolution

A priori, we might suppose that divergent evolution of sequences from the same ancestor would give rise to a distribution of sequence identity scores with a peak value, D, at some probably low value, e.g. $D < 30\%$, and a smooth relatively flat tail for high values (Fig. 4). In the case of convergent evolution, where two unrelated sequences evolve to the same structure, we would expect a sharp Gaussian distribution with a peak value, C, at very low identity, e.g. $C < 10\%$ (Fig. 4).

Observed distribution for structural homologues: one peak at ~8–9%

The distribution of sequence identity scores for structural homologues (Fig. 3) has three distinct regions: a large, approximately Gaussian peak centred at ~8–9%; many smaller sharp peaks between 15 and 95%; and a large peak near 100%. The last peak may arise from mutants engineered to facilitate structure determination. The second region can be explained by 'incoherent noise' peaks arising from uneven sampling and the still relatively small size of the current PDB (see below). The peak in the first region seemed to be incompatible with the hypothesis that convergent and divergent evolution yielded two different Gaussian distributions (around C and D); the observed peak occurs at very low average identity (~8–9%) and is remarkably symmetrical (Fig. 3, left panel). The peak is also very similar to the distribution of random sequence

Figure 4



Hypothetical distribution of pairwise sequence identity for two evolutionary events. Firstly, protein pairs that converged from a different ancestor to similar structures (grey line; peak at *C*). Secondly, proteins that diverged from a common ancestor maintaining a similar structure (black line; peak at *D*). The dashed line indicates that it is not clear *a priori* which relation to expect between the divergent peak and its tail at high levels of sequence identity.

identities with a peak value, *R*, at ~5.6% (Fig. 3, left panel). Qualitatively similar results were obtained when using sequence similarity instead of sequence identity [21].

Have divergent and convergent evolution reached a similar equilibrium?

Three scenarios could have generated the observed distribution for similar structures with vanishing pairwise sequence identity (<15%) as a superposition of two separate events (Fig. 3).

1. Divergent evolution has not reached down to very low levels of sequence identity; the observed distribution for remote homologues is entirely dominated by pairs that converged from different ancestors to adopt similar structures.

2. Convergent evolution is negligible; all observed pairs have originated from divergence to very low levels of sequence identity.

3. Divergent and convergent evolution have reached similar equilibrium distributions.

Divergent evolution not underrepresented for remote structural homologues

The underrepresentation of pairs that have diverged from a common ancestor may have been caused by the particular definition of structural similarity (the more 'relaxed' the definition, the more likely that even functionally unrelated proteins could be deemed 'structurally similar'). However, various different criteria yielded qualitatively the same result: a single Gaussian distribution peaking at ~8–9% sequence identity explained the observed data best (Fig. 5).

Similar results for alignment of inter- and intra-kingdom data

The one peak at ~8–9% sequence identity (Fig. 2) may have arisen from alignments between proteins from the same terrestrial kingdom (e.g. prokaryotes with prokaryotes, or eukaryotes with eukaryotes) whereas the several peaks above 30% sequence identity may have resulted from alignments between proteins from different terrestrial kingdoms (e.g. prokaryotes with eukaryotes). Compiling the distributions separately for all prokaryote–prokaryote, all eukaryote–eukaryote, and all eukaryote–prokaryote alignments did not confirm this suspicion. Instead, for all inter- and intra-kingdom alignments, the distributions appeared qualitatively similar to the one for all protein pairs (Fig. 2).

Most close structural homologues have less than 45% pairwise sequence identity

Sequence alignments of all proteins from the four entire genomes against SWISS-PROT and TREMBL databases [29] yielded several clear results.

1. The coherent peak near 100% (Fig. 3, right panel) is not present for any of the genomes (Fig. 6).

2. The various smaller peaks between 40 and 80% in the distribution of structural alignments (Fig. 3, right panel) are not coherently observed in the four genomes (Fig. 6).

3. The majority of close structural homologues (>30% pairwise sequence identity) for all four genomes had 30–42% pairwise sequence identity (data not shown).

Discussion

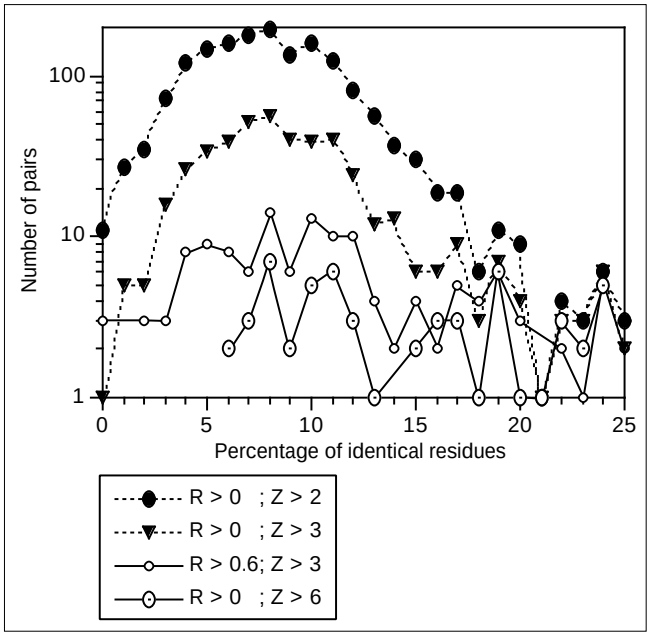
How stable is protein structure with respect to sequence changes?

Most pairs of similar structures have sequence identity as low as expected from randomly related sequences (Fig. 3, left panel). This does not imply that sequence changes were random but that to us — as observers of the record of evolutionary history — the sequence variations look random.

How many 'anchor' residues define a structure?

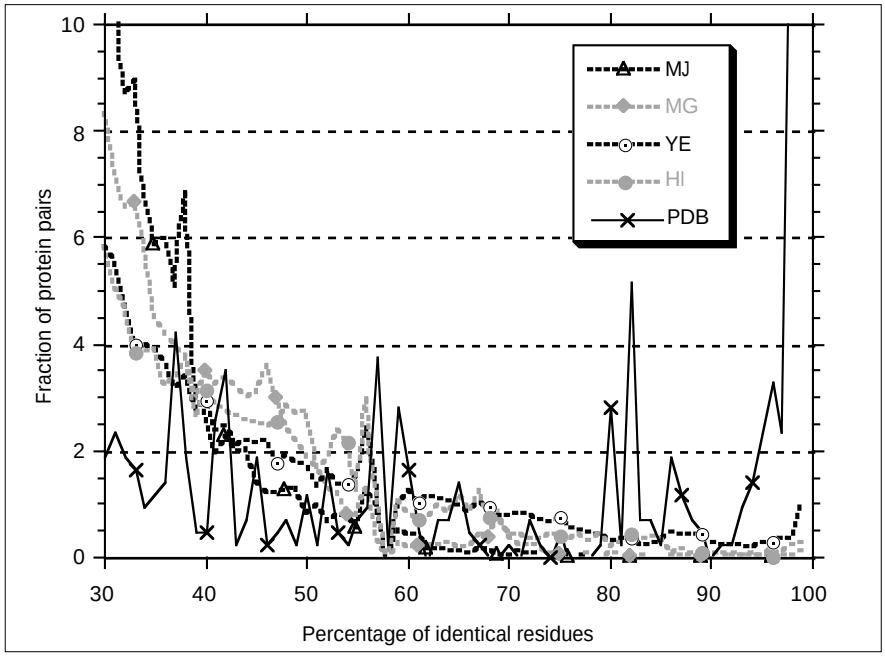
The average percentage 'anchor' residues, i.e. residues that are crucial for protein structure and function, is not given by the average of the observed distribution. Instead, the relevant

Figure 5



Distribution of pairwise sequence identity for remote structural homologues using different thresholds for structural similarity. Different thresholds were generated by excluding all pairs in the FSSP database [22] below a given threshold in the DALI Z-score (Z-DALI < 2, 3, and 6) [23], and for which the alignment covered too small a percentage of the aligned proteins (R < 0, 0.6). Note the logarithmic scale of the vertical axis; in logarithmic scale, a Gaussian curve becomes a parabola. By tuning the threshold, rather different data sets were generated. At the most stringent cut-off value (Z < 6), too few examples for statistical analysis were found in the PDB. However, the plots suggest that the details of the shape of the function displayed in Figure 3 were independent of the particular choice of the cut-off for structural similarity.

Figure 6



Distribution of pairwise sequence identity for all close structural homologues between four entire genomes and the SWISS-PROT database. The results from the analysis of structural pairs are also plotted. Counts are normalized to percentages by the sum over all pairs for each genome (respecting the structural pairs). Values between the structural alignments (PDB) and the sequence alignments (genomes) were not strictly comparable. However, the main trend is clear: the peaks in the PDB distribution at around 80 and 100% sequence identity arise from bias in the selection of structures in the PDB. MJ, *Methanococcus jannaschii* ; MG, *Mycoplasma genitalium*; YE, *Saccharomyces cerevisiae* (yeast); HI, *Haemophilius influenzae*.

number is the difference between the peak observed for structural homologues (O = 8%; Fig. 3), and the peak for random alignments (R). Thus, on average only 3–4% of the residues anchor a protein structure ($O - R$; Fig. 3).

Did evolution have enough time to reach an equilibrium?

Most remote structural homologues have <15% pairwise sequence identity (Fig. 3), and most close structural homologues have <45% pairwise sequence identity (Fig. 5). This implies that the rate of creation of new structures is much slower than the drift towards the mean sequence identity (D). Furthermore, the symmetric shape of the distribution at low sequence identity suggests that four billion years of evolution was sufficient to reach an equilibrium between these two processes.

Can we distinguish between convergent and divergent evolution?

Naively, we may have supposed that the level of pairwise sequence identity for remote homologues can be used to distinguish between convergent and divergent evolution. However, the results presented here suggest that convergent and divergent evolution may have quite similar equilibrium states (difference between divergent and convergent average, $D - C$, quite small; Fig. 3), and hence, in the remote homology region (<15%), it is difficult to distinguish between the two effects.

How will the distribution look for all globular proteins?

Assuming that the three terrestrial kingdoms (eukaryote, prokaryote and archean) would result in separated clusters, the observed distributions could also be attributed to such clusters. The relation between the major peak, ~8–10% sequence identity, and the minor peaks, ~60 and 80% (Fig. 2), would then reflect merely the distribution of organisms that are the sources for the protein sequences in the PDB. Furthermore, the suggestion from Figure 2 that most structural homologues have <15% pairwise sequence identity may also be a result of the particular distribution of organisms in the PDB. However, restricting the distributions to alignments from the same kingdom and to alignments between two different kingdoms did not qualitatively alter the distributions (Fig. 4). The fact that very different data sets produced similar results suggests that the distribution for all globular proteins would look similar to the one observed today. However, the data sets are still too small to allow drawing firm conclusions.

Trivial or surprise?

In presenting this analysis at the Copenhagen meeting and elsewhere, most experts have expressed surprise at the low value of the average pairwise sequence identity. Clearly, then, the distributions shown in Figures 1 and 3 contain an important lesson in advancing our understanding of the evolution of proteins.

Acknowledgements

I thank Sean O'Donoghue (EMBL, Heidelberg) for crucial encouragement, many discussions, extremely valuable remarks, and proofreading. Thanks also to Chris Sander (EBI, Hinxton) for his intellectual and financial support. Furthermore, thanks to the extremely constructive comments from one of the anonymous referees. In general, thanks to all those who deposit information in public databases, and to those who carry the burden of maintaining these valuable evolutionary records.

References

1. Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.J., Davies, D.R. & Phillips, D.C. (1960). Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature* **185**, 422–427.
2. Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, G., Will, G. & North, A.T. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature* **185**, 416–422.
3. Bernstein, F.C., *et al.,* & Tasumi, M. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
4. Lattman, E.E. (1994). Protein crystallography for all. *Proteins* **18**, 103–106.
5. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science* **283**, 595–602.
6. Gerstein, M. & Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. In *Fourth International Conference on Intelligent Systems for Molecular Biology*. (States, D., Agarwal, P., Gaasterland, T., Hunter, L. and Smith, R.F., eds.), pp 59–67, AAAI Press, St. Louis, MO, USA.
7. Holm, L. & Sander, C. (1996). Alignment of three-dimensional protein structures: network server for database searching. *Meth. Enzymol.* **266**, 653–662.
8. Orengo, C.A. & Taylor, W.R. (1996). SSAP: sequential structure alignment program for protein structure comparison. *Meth. Enzymol.* **266**, 617–635.
9. Zu-Kang, F. & Sippl, M.J. (1996). Optimum superimposition of protein structures: ambiguities and implications. *Fold. Des.* **1**, 123–132.
10. Luo, Y., Lai, L., Xu, X. & Tang, Y. (1993). Defining topological equivalences in protein structures by means of a dynamic programming algorithm. *Protein Eng.* **6**, 373–376.
11. Mizuguchi, K. & Go, N. (1995). Seeking significance in three-dimensional protein structure comparisons. *Curr. Opin. Struct. Biol.* **5**, 377–382.
12. Hilbert, M., Böhm, G. & Jaenicke, R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins* **17**, 138–151.
13. Crippen, G.M. & Maiorov, V.N. (1995). How many protein folding motifs are there? *J. Mol. Biol.* **252**, 144–151.
14. Alexandrov, N.N., Takahashi, K. & Go, N. (1992). Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.* **225**, 5–9.
15. Overington, J.P. (1992). Comparison of three-dimensional structures of homologous proteins. *Curr. Opin. Struct. Biol.* **2**, 394–401.
16. May, A.C.W. & Johnson, M.S. (1994). Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. *Protein Eng.* **7**, 475–485.
17. Chothia, C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
18. Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* **9**, 56–68.
19. Flahetry, K.M., McKay, D.B., Kabsch, W. & Holmes, K.C. (1991). Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70K heat-shock cognate protein. *Proc. Natl. Acad. Sci. USA* **88**, 5041–5045.
20. Holmes, K.C., Sander, C. & Valencia, A. (1993). A new ATP-binding fold in actin, hexokinase and Hsc70. *Trends Cell Biol.* **3**, 53–59.
21. Rost, B., O'Donoghue, S. & Sander, C. (1996). Protein structures evolve at random, almost. World Wide Web URL: http://www.embl-heidelberg.de/~rost/Papers/PreEvolution96.html
22. Holm, L. & Sander, C. (1996). The FSSP database: fold classification based on structure–structure alignment of proteins. *Nucleic Acids Res.* **24**, 206–210.
23. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.
24. Doolittle, R.F. (1986). *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences.* University Science Books, Mill Valley, California.
25. Fleischmann, R.D., *et al.,* & Venter, J.C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
26. Goffeau, A., *et al.* & Oliver, S.G. (1996). Life with 6000 genes. *Science* **274**, 546–567.
27. Fraser, C.M., *et al.,* & Venter, J.C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.
28. Bult, C.J., *et al.,* & Geoghagen, N.S.M. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073.
29. Bairoch, A. & Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* **24**, 21–25.
30. Schneider, R. & Sander, C. (1996). The HSSP database of protein structure–sequence alignments. *Nucleic Acids Res.* **24**, 201–205.